Prospective Validation of Text Categorization Filters for Identifying High-Quality, Content-Specific Articles in MEDLINE.

Y. Aphinyanaphongs, M.S., C.F. Aliferis M.D., Ph.D.
Department of Biomedical Informatics, Vanderbilt University, Nashville, TN

Abstract

Finding high quality articles is increasingly difficult with the exponential growth of the medical literature. This growth requires new methods to identify high quality articles. In prior work, we introduced a machine learning method to identify high quality MEDLINE documents in internal medicine. The performance of the original filter models built with this corpus on years outside 1998-2000 was not assessed directly. Validating the performance of the original filter models on current corpora is crucial to validate them for use in current years, to verify that the model fitting and model error estimation procedures do not over-fit the models, and to validate consistency of the chosen ACPJ gold standard (i.e., that ACPJ editorial policies and criteria are stable over time). Our prospective validation results indicated that in the categories of treatment, diagnosis, prognosis, and etiology, the original machine learning filter models built from the 1998-2000 corpora maintained their discriminatory performance of 0.95, 0.97, 0.94, and 0.94 area under the curve in each respective category when applied to a 2005 corpus. The ACPJ is a stable, reliable gold standard and the machine learning methodology provides robust models and model performance estimates. Machine learning filter models built with 1998-2000 corpora can be applied to identify high quality articles in recent years.

Introduction

The purpose of a query filter is to identify medical articles that meet certain criteria (e.g., related to quality, impact, or content). Recent approaches have utilized machine learning or semi-manually constructed Boolean query based filters to pre-select articles that meet quality and content criteria [1-5]. These filters had good discriminatory performance when evaluated using cross-validation techniques [6].

Both machine learning and Boolean filters can perform much worse than expected when applied to other corpora because of two main reasons: First, it is possible for filters to be over-fitted, and second, the examples that were used to train the original filters may have a different distribution than the documents on which the filters are eventually applied [7].

Computational Learning Theory suggests that over-fitting typically occurs when filter developers fit model parameters using the training data and then estimate the future performance of the model on the same data, or when very complex models are pursued, relative to the classification function's intrinsic complexity especially in small sample learning settings (i.e., the complexity of the models considered is not tempered by the available sample and the difficulty of the learning task) [8]. Sound data modelling principles in order to avoid over-fitting include: (a) choosing model complexity and parameters that minimize both error in the training data and complexity of the model class employed; (b) estimating future (generalization) error in portions of the data reserved especially for that purpose (i.e., they are not used to fit the model) [9].

With regards to filter failure because of non-representative samples, this may occur because of small samples or very rare positive examples even if the total sample is large. In addition, non i.i.d. (independently sampled and identically distributed) sampling from the general population of documents. may lead to divergence of the training document set distribution from the application document set distribution. A particularly worrisome reason for violation of i.i.d. sampling in our context is if the gold standard for document labelling is not stable over time. For example, if the editorial policies of the ACP Journal Club changes over time, a filter built with an older editorial policy may exhibit worse performance for documents characterized as high-quality according to a more recent and thus revised editorial policy.

In this study, we address these points of failure for both Machine Learning (i.e., our own) and Boolean/semi-manual (i.e., Pubmed/Haynes et al's Clinical Query (CQ)) filters. We explore the extent of over-fitting or changes in the characteristics of the data by evaluating classification performance on articles collected independently of the original corpus. We built a machine learning filter model using a training corpus collected in one year, evaluated its performance on a prospective testing corpus collected in another year, and in the same prospective corpus, compared the machine learning filter models to the CQ filters [10] of Pubmed[1]. Thus, we have *two main hypotheses*. First, machine

[1] The CQF filters are literally Boolean combinations of terms applied to a corpus. The machine learning filter models, in contrast, are not Boolean based. The machine learning filters are statistical models using all the terms in the training corpus.

learning filter models built from an original corpus collected from 1998 to 2000 are able to identify high quality articles in an independently collected 2005 corpus and perform as well as estimated performance measures using cross-validation on the original corpus. Second, machine learning filter models retain their performance edge over the corresponding CQ filters in the 2005 corpus.

Methods

Definitions

At the core of our efforts lie the selection of a rigorous quality, content gold standard and the creation of a document collection that captures this gold standard. The ACP journal club is a highly-rated meta-publication [11]. Every month experts review the best journals in internal medicine and select the best articles according to specific selection criteria in the article class areas of: *treatment, diagnosis, etiology, prognosis, quality improvement, clinical prediction guide*, and *economics*. Selected articles are further subdivided into articles that are cited and abstracted by the ACP because of their clinical importance, and those that are only cited because they meet all the selection criteria but may not pertain to vitally important clinical areas. Every article is subjected to rigorous review for inclusion [11]. By using articles abstracted and cited by the ACP as our gold standard, we capitalize on an existing high quality review.

Corpus Construction

We constructed corpora in the treatment, etiology, diagnosis, and prognosis categories spanning the time periods of July 1998 – August 1999, July 1998 – August 2000, and March 11, 2005 – August 31, 2005. From [1], we reused the two corpora built from the first two periods. For each corpus, we started with 49 journals, selected the respective time period, and collected all articles with abstracts published by these journals. We then reviewed the ACP Journal Club for at least 18 months after the specified time period for each corpus, and labeled as positive any article that was cited/ abstracted by the Journal Club in the time period. The first corpus spanning July 1998 – August 1999 resulted in a positive/negative article distribution of 379/ 15,407 articles in treatment, and 205/ 15,581 articles in etiology. The second corpus spanning July 1998 – August 2000 resulted in a positive/negative article distribution of 74/34,864 articles in prognosis, and 102/34,836 articles in diagnosis. Refer to [1] for additional details and motivations for these constructed corpora.

We constructed the third corpus for the prospective analysis from March 11, 2005-August 31, 2005. We built the third corpus using the electronic citations available from the ACP Journal club at http://www.acpjc.org. Both articles cited and abstracted and articles cited only were available in both the print and electronic versions of the journal club. As of July 2005, the electronic version included an expanded list of articles cited only available at http://www.acpjc.org/Content/oan which we included in the independent dataset.

We covered available electronic citations in the Journal Club from July/Aug 2005 to Jan/Feb 2006 in 41 journals selected for their overlap with the 1998-2000 49 journals[2]. Because the time frame covered by the Journal Club varied from month to month, we selected 3/11/2005 as the start time period for this third corpus by averaging the earliest citation given in each journal, and the end time period of 8/31/2005 by averaging the latest citation given in each of the 41 selected journals. If no article occurs in a given journal, a date is not included in the average. Thus we selected all articles with abstracts published in 41 journals from 3/11/2005 to 8/31/2005 and identified articles cited in this time period by the ACP Journal club as positive and all others were identified as negative. This procedure resulted in a positive/negative article distribution of 351/6,921, 47/7,601, 30/7,618, 23/7,625 in treatment, etiology, prognosis, and diagnosis respectively.

All original articles as Pubmed citations (i.e. abstracts, not full text) were downloaded with the esearch and efetch utilities available from Pubmed [12]. Each search was limited to the title of one of the journals, set to only retrieve articles during the publication period, and with the "only items with abstracts" checkbox marked. A custom parser extracted PubmedID, title, journal, abstract, publication type, and MeSH terms from the XML efetch downloads.

Article Preparation

The conversion of documents to a format suitable for the machine learning algorithm followed the procedures in [1]. The articles in the ACPJ selected journals were cross-referenced in PubMed, and the title, abstract, journal, publication type, and MeSH terms were extracted. We created two representations for each document: one for the machine learning algorithm, and one for the CQ filters.

For the machine learning algorithm, we represented each document as a set of terms for the learning algorithms [13]. We additionally stemmed each term [14], removed "stopword" terms [15], and removed any terms occurring in fewer than 5 documents. Very infrequent terms are difficult to

[2] Journal lists for both corpora are available from the authors.

assess statistically and may affect negatively the generalization of the classification models. Terms were further encoded as weighted features using a log frequency with redundancy scheme [16].

For the CQ filters, we represented each document as a set of terms. Words were not stemmed, but "stopwords" and infrequent terms (occurring in < 5 documents) were removed.

Statistical and Machine Learning Methods

Support Vector Machines (SVMs)

In our experiments, we employed Support Vector Machine (SVM) classification algorithms. The SVM's calculate maximal margin hyperplane(s) separating two or more classes of the data. To accomplish this, the data are mapped to a higher dimensional space by means of a kernel function, where a separating hyperplane is found by solving a constrained quadratic optimization problem [17]. SVMs have had superior text classification performance compared to other methods [1, 18], and this motivated our use of them. We used an SVM classifier implemented in libSVM v2.8 [19] with a polynomial kernel. We optimized the SVM penalty parameter C over the range {0.1, 1, 2} with imbalanced costs applied to each class proportional to the priors in the data [20], and degree d of the polynomial kernel over the range {1, 2}. Since theoretical literature on domain characteristics as it relates to optimal parameter selection is not yet developed, the ranges of costs and degrees for optimization were chosen based on previous empirical studies [1, 18]. Different combinations of costs and degrees were exhaustively evaluated by cross-validation.

Clinical Query Filters

The CQ filters are Boolean queries optimized separately for sensitivity, specificity, and accuracy [10]. We applied the exact queries for optimized sensitivity and specificity cited in Pubmed and recently updated with a year 2000 corpus to the text categorization task [2-4].

Estimating Model Performance

We used 5-fold cross-validation that avoids over-fitting to estimate the performance of the learning algorithms [6]. This choice for n provided sufficient high-quality positive samples for training in each category and provided sufficient article samples for the classifiers to learn the models. The cross-validation procedure first divided the data randomly into 5 non-overlapping subsets of documents where the proportion of positive and negative documents in the full dataset is preserved for each subset. Next, the following was repeated 5 times: we used one subset of documents for testing (the "original testing set") and the remaining four subsets for training (the "original training set") of the classifier. The average performance over 5 original testing sets is reported.

In order to optimize parameters of the SVM algorithms, we used another "nested" loop of cross-validation by further splitting each of the 5 original training sets into smaller training sets and validation sets. For each combination of learner parameters, we obtained cross-validation performance and selected the best performing parameters inside this inner loop of cross-validation. We next built a model with the best parameters on the original training set and applied this model to the original testing set. Details about the "nested cross-validation" procedure can be found in [7, 21]. Notice that the final performance estimate obtained by this procedure will be unbiased because each original testing set is used only once to estimate performance of a single model that was built by using training data exclusively.

Applying Filters to Prospective Corpora

We built final machine learning filter models in each category using the 1998-1999 and 1998-2000 corpora and then applied both the final machine learning filter models and the CQ filters to the prospective 2005 corpus. We built the final machine learning filter models by selecting best performing parameters (i.e. cost and degree) and applying these parameters to build final models in each category using all the data. Best parameters were selected by first, dividing the data into 5 non-overlapping subsets preserving positive/ negative proportions. For each set of parameters, we estimated performance using cross-validation over the 5 folds. Average performance across all folds with each set of parameters was recorded. We selected the parameters that built the best performing filter model, and used these parameters to build a final machine learning filter model for each category using all the data.

Comparing CQ Filters to Learning Models

We compared the sensitivity and specificity of the machine learning filter models with the sensitivity and specificity of the respective optimized Boolean CQ filter [10]. The CQ filters return articles with the query terms present, whereas the learning algorithms return a score. To make the comparison, in each fold, we fixed the sensitivity value returned by the sensitivity-optimized CQ filter and varied the threshold for the scored articles until the sensitivity was matched. We report the fixed sensitivity, corresponding specificity, and precision. The same procedure was run for the specificity returned by the optimized specificity CQ filter.

Results

Area under the curve analysis

We built machine learning filter models for treatment, etiology, prognosis, and diagnosis categories using the 1998-1999 and 1998-2000 corpora. In Table 1, we report the cross-validation area under the ROC curve for the 1998-1999 and 1998-2000 built machine learning filter models, and area under the ROC curve performance when the machine learning filter models were applied to the entire 2005 corpora in the 4 categories.

Table 1: Top row is cross-validation estimated area under the curve for optimal 1998-1999 and 1998-2000 models. Bottom row is area under the curve for the optimal models applied to 2005 corpora (*no cross-validation applied*). Treat – treatment, Diag – diagnosis, Prog-prognosis, Etio – etiology. ± - is the range of AUC estimates across the 5 folds.

	Treat	**Diag**	**Prog**	**Etio**
X-Val AUC	0.97± .02	0.99 ± .02	0.95± .02	0.95 ± .01
2005	0.95	0.97	0.94	0.94

The optimal machine learning filter models built using the 1998-1999 and 1998-2000 corpora and applied to the 2005 corpora had performances within the range of estimates of each fold in each cross-validation set. The optimal machine learning filter models were able to discriminate high quality articles from other non-high-quality articles in the 2005 corpora.

Comparison to CQ filters

We applied the CQ filters of Pubmed to the entire 2005 corpora and reported their corresponding sensitivity and specificities in Table 2. In all 4 categories, the CQ filters performed well. The support vector machine outperforms the CQ filters in sensitivity, specificity, and precision at fixed sensitivity and specificity levels.

The specificity and sensitivity optimized prognosis CQ filters and specificity optimized etiology CQ filters have lower sensitivity and specificity than previously reported results. The sensitivity optimized prognosis CQ filter (90.0% as reported in [3] vs. 80.0% in the current study), specificity optimized prognosis CQ filter (94.1% as reported in [3] vs. 76.8% in the current study), and the specificity optimized etiology CQ filter (94.9% as reported in [2] vs. 83.9% in the current study) do not perform as expected. Further investigation is necessary to determine the cause of this performance discrepancy and possible solutions.

Discussion

These experiments addressed a pertinent and important question for using a filter to identify articles in a corpus. If we built machine learning or apply semi-manually constructed Boolean-based CQ filters using a corpus from a different time period, can we reliably apply these filters to current corpora and identify the high quality articles.

Our results showed that we can identify articles in this 2005 corpus using CQ filters or machine learning filter models. The optimized machine learning filter models built with the 1998-1999 and 1998-2000 corpora from [1] do generalize as estimated by the cross-validation procedure and were able to identify high quality articles accurately in a 2005 corpora as measured by area under the curve. The CQ filters of Pubmed were also able to identify high quality

Table 2 – Optimized Support Vector Machine (SVM) compared to Clinical Query Filters fixed at optimal sensitivity and specificity. All values are calculated using the entire 2005 corpora.

Category	Optimized For	Method	Sensitivity	Specificity	Precision
Treatment	Sensitivity	Query Filters	**0.980**	0.710	0.147
		SVM		0.888	0.305
	Specificity	Query Filters	0.803	**0.913**	0.318
		SVM	0.948		0.349
Etiology	Sensitivity	Query Filters	**0.979**	0.435	0.010
		SVM		0.753	0.024
	Specificity	Query Filters	0.681	**0.839**	0.025
		SVM	0.936		0.035
Diagnosis	Sensitivity	Query Filters	**0.956**	0.682	0.01
		SVM		0.884	0.02
	Specificity	Query Filters	0.652	**0.972**	0.07
		SVM	0.821		0.08
Prognosis	Sensitivity	Query Filters	**0.800**	0.707	0.011
		SVM		0.874	0.024
	Specificity	Query Filters	0.800	**0.768**	0.013
		SVM	1.00		0.017

articles. As anticipated by [1], the optimized machine learning filter models generalize well and had superior ability over the optimized CQ filters to identify quality articles in the 2005 corpus.

These results also validate the optimization methods used to build the machine learning filter models and the consistent editorial policies of the ACP Journal Club. The ability of the 1998-1999 and 1998-2000 corpora based machine learning filter models to identify high quality articles in the 2005 corpus imply that the procedure to optimize the machine learning filter model (through cross-validation) is valid and creates robust models and model performance estimates.

Furthermore, the ACP Journal Club is a consistent, stable gold standard. The 1998-1999 and 1998-2000 based corpora machine learning filter models discriminatory power to identify high quality articles succeeds due to consistent article selection in the original and prospective corpora. The machine learning filter models prediction of high quality articles in the 2005 corpora imply that the methodologic criteria for high quality articles has not changed over time, and we may reliably apply these machine learning filter models in current years.

The true purpose of any filter is to identify high quality articles in later corpora. This paper is a step to validating filters for medical information retrieval. Coupled with our previous work [1], we are establishing a foundation for usage of these filters.

In current work, we are systematically evaluating these filters in answering "real-life" clinical questions. As a first step, we have built a prototype at www.ebmsearch.org. How well these filters can assist expert reviewers and their generalization to other categories and domains are open questions that we have experiments underway to answer.

Acknowledgements

The first author acknowledges support from NLM grant LM007948-02. The second author acknowledges support from grant LM007948-01.

References

1. Aphinyanaphongs Y, Tsamardinos I, Statnikov A, Hardin D, Aliferis CF. Text Categorization Models for High Quality Article Retrieval in Internal Medicine. J Amer Med Inform Assoc. 2005;12(2):207-216.
2. Wilczynski N, Haynes B. Developing Optimal Search Strategies for Detecting Clinically Sound Causation Studies in MEDLINE. In: Proc AMIA Symposium; 2003; Washington DC; p. 719-23.
3. Wilczynski N, Haynes B. Optimal Search Strategies for Detecting Clinically Sound Prognostic Studies in EMBASE. J Amer Med Inform Assoc. Jul/Aug 2005;12(4):481-485.
4. Haynes B, Wilczynski N. Optimal Search Strategies for retrieving scientifically strong studies of diagnosis from MEDLINE: an analytical survery. BMJ 2004.
5. Wilczynski N, Haynes B. Robustness of Empirical Search Strategies for Clinical Content. In: AMIA; 2002.
6. Weiss S, Kulikowski CA. Computer Systems that Learn. San Mateo, CA: USA M. Kauffman; 1991.
7. Scheffer T. Error estimation and model selection. Technischen Universit at Berlin; 1999.
8. Kearns M, Umesh V. An Introduction to Computational Learning Theory: MIT Press; 1994.
9. Aliferis CF, Statnikov A, Tsamardinos I. Challenges in the Analysis of Mass-Throughput Data. Cancer Informatics. To appear 2006.
10. (Accessed: 03-13-2006), http://www.ncbi.nlm.nih.gov/entrez/query/static/clinical.html.
11. ACP_Journal. Purpose and Procedure. ACP Journal 1999;131(1):A-15 - A-16.
12. PubMed. (Accessed: 3-06-2006), http://www.ncbi.nlm.nih.gov/PubMed/.
13. Salton G, Buckley C. Term weighting approaches in automatic retrieval. Information Processing and Management 1988;24(5):513-523.
14. Porter MF. An algorithm for suffix stripping. Program 1980;14(3):130-137.
15. MEDLINE Stopwords. (Accessed: 3-13-2006), http://biolib.princeton.edu/instruct/MedSW.html.
16. Leopold E, Kindermann J. Text Categorization with Support Vector Machines. How to Represent Texts In Input Space? Machine Learning 2002;46:423-444.
17. Vapnik V. Statistical Learning Theory. New York: Wiley; 1998.
18. Joachims T. Text Categorization With SVMs: Learning With Many Relevant Features. In: Proceedings of the 10th European Conference On Machine Learning; 1998: Springer-Velag.
19. LIBSVM: a library for support vector machines. (Accessed: 3-13-2006), http://www.csie.ntu.edu.tw/~cjlin/libsvm.
20. Morik K, Brockhausen P, Joachims T. Combining statistical learning with a knowledge-based approach - A case study in intensive care monitoring. In: Proc. 16th Int'l Conf. on Machine Learning (ICML-99); 1999.
21. Dudoit S, Van Der Laan MJ. Asymptotics of cross-validated risk estimation in model selection and performance assessment. Working Paper: U.C. Berkeley Division of Biostatistics; 2003 February 5. Report No.: 126.